Assignment #2

Jingye Wang

September 19, 2016

Gelman & Hill 3.4

```
# data imput
setwd("~/Dropbox/WUSTL third/Multilevel Modeling for Quantitative Research/assignmen
t")
library(foreign)
library(car)
library(arm)
library(ggplot2)
library(MASS)
library(visreg)
iq_data <- read.dta ("child.iq.dta")
```

Part A

Fit a regression of child test scores on mother's age, display the data and fitted model, check assumptions, and interpret the slope coefficient. When do you recommend mothers should give birth? What are you assuming in making these recommendations?

Fit a regression model

```
reg_1 <- lm(ppvt ~ momage, data=iq_data)
display(reg_1)
```

```
lm(formula = ppvt ~ momage, data = iq_data)
            coef.est coef.se
(Intercept) 67.78     8.69
momage       0.84     0.38
---
n = 400, k = 2
residual sd = 20.34, R-Squared = 0.01
```

Check assumptions

1.Multicollinearity

As this model has only one predictor (momage), this assumption does not apply here.

2.Independence of errors

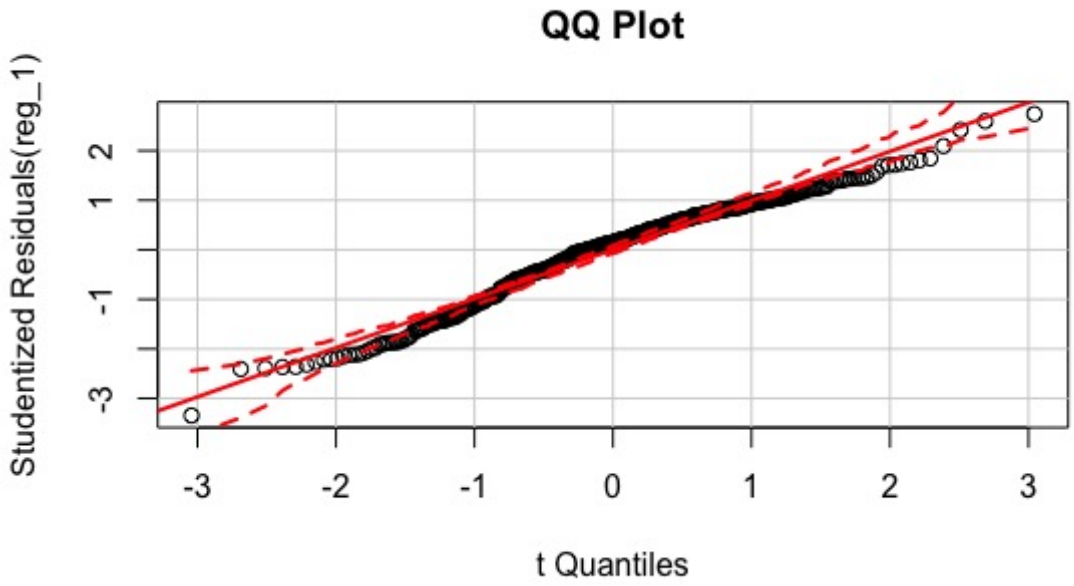
```
durbinWatsonTest(reg_1)
```

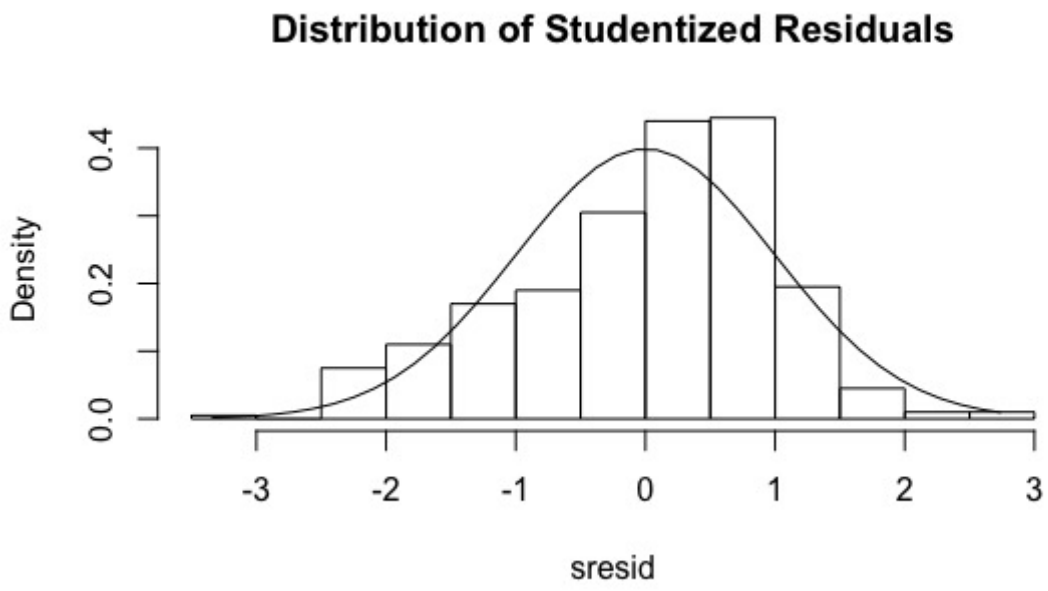
```
 lag Autocorrelation D-W Statistic p-value
   1     -0.01389737      2.020368   0.824
 Alternative hypothesis: rho != 0
```

The Durbin-Watson test was conducted and the statistic was 2.020368 close to 2 and neither less than 1 or larger than 3, thereby meeting this assumption.

3.Normality of errors

```
qqPlot(reg_1, main="QQ Plot")
sresid <- studres(reg_1)
hist(sresid, freq=FALSE,
     main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```





The QQplot and the histogram of studentized residuals strategy was used. As shown in the figure above the studentized residuals were normally distributed. This assumption was met.

4.Homoscedasticity

non-constant error variance test

```
ncvTest(reg_1)
```

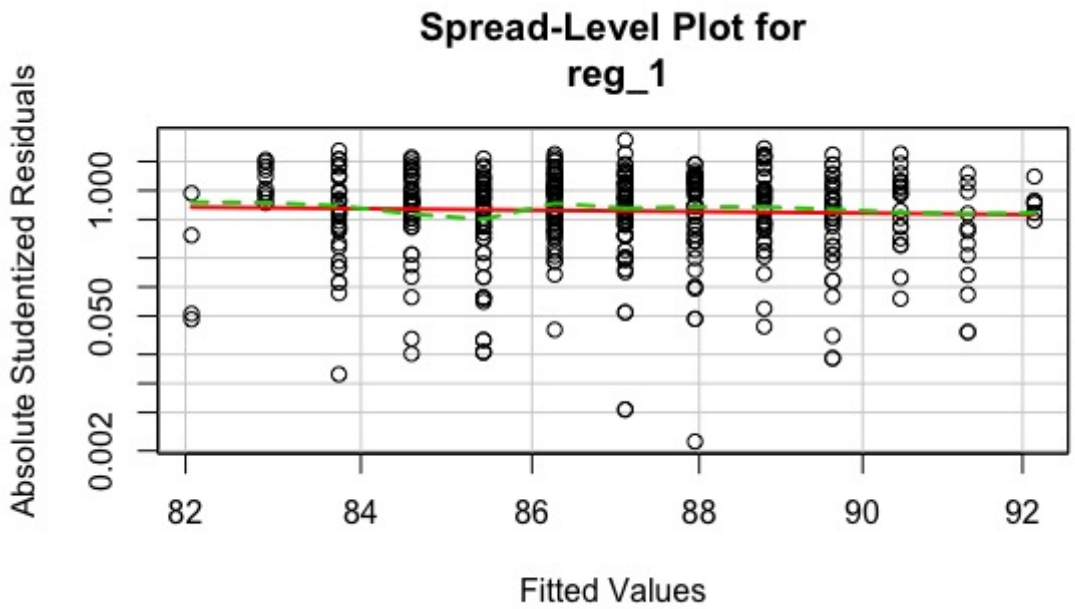
```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.3873767    Df = 1     p = 0.5336815
```

As shown by the non-constant error variance test, the results of the Chi-square test was non-significant [χ2(1)=0.3873767, P= 0.5336815]. This assumption was met.

5.Linearity

```
spreadLevelPlot(reg_1)
```

```
Suggested power transformation:  2.565026
```



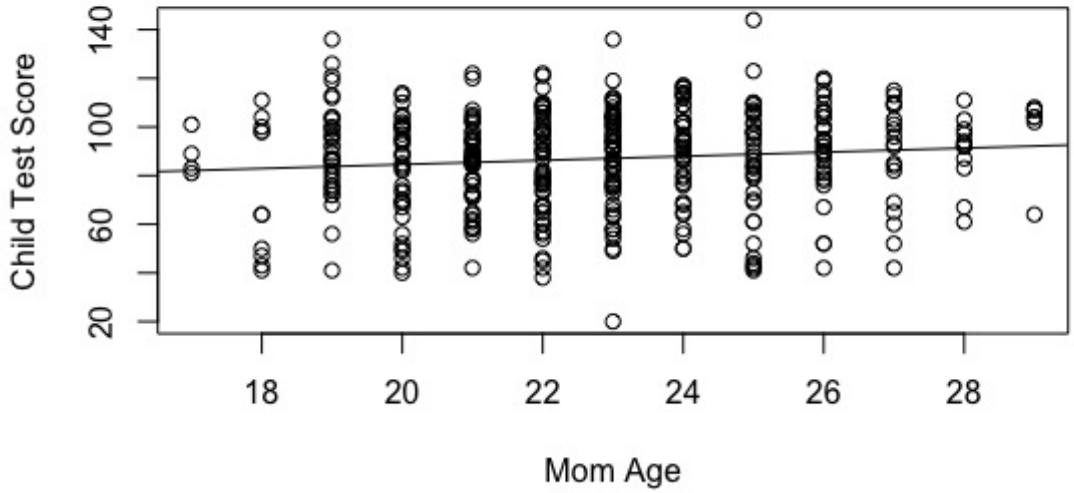
Based on the Figure, points spread evenly around the line. This assumption was met.

interpret the slope coefficient

When the mothers' age increases by 1, the Child Test Score increases by 0.8403.

recommendations

```
plot(iq_data$momage, iq_data$ppvt, xlab="Mom Age", ylab="Child Test Score")
abline(reg_1)
```



Based on the above graph and the results of the regression model, I would suggest mothers should give birth on their late 20s. However, mother's age was not a strong predictor for child test score and the adjusted R-squared for this model was only 0.009743. So we could not just really rely on this model to interpret the relationship.

Part B

Repeat this for a regression that further includes mother's education, interpreting both slope coefficients in this model. Have your conclusions about the timing of birth changed?

Fit model

```
reg_2 <- lm(ppvt ~ momage + educ_cat, data=iq_data)
display(reg_2)
```

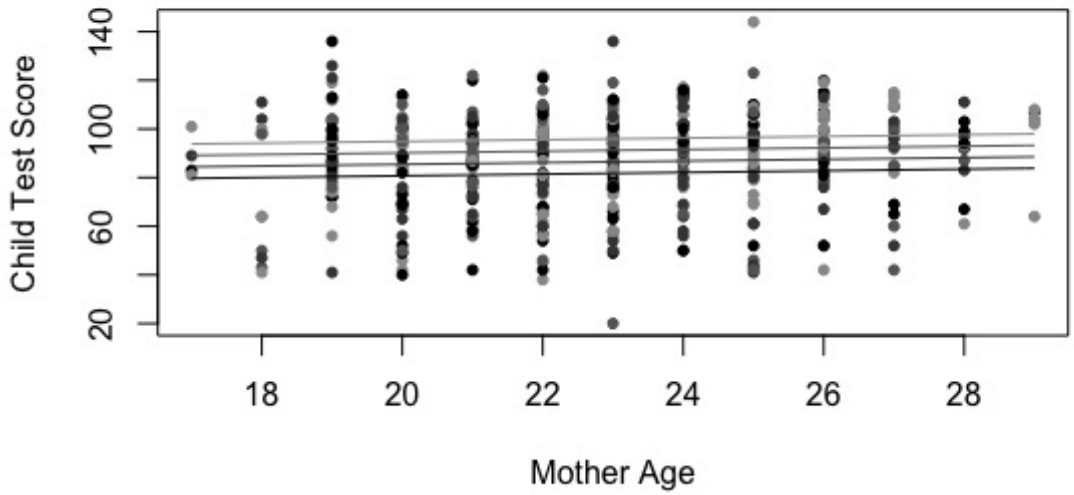
```
lm(formula = ppvt ~ momage + educ_cat, data = iq_data)
            coef.est coef.se
(Intercept) 69.16     8.57
momage       0.34     0.40
educ_cat     4.71     1.32
---
n = 400, k = 3
residual sd = 20.05, R-Squared = 0.04
```

Interpreting slope coefficients

When the mothers' age increases by 1, the Child Test Score increases by 0.3433. This association is not statistically significant (t=0.3981, P=0.389003). When the mothers' education level increases by 1, the Child Test Score increases by 4.7114. This association is statistically significant (t=3.579, P=0.000388).

Conclusions

```
colours = c('#000000','#444444','#666666','#999999')
plot(iq_data$momage, iq_data$ppvt, xlab="Mother Age", ylab="Child Test Score",
     col=colours, pch=20)
for (i in 1:4) {
  curve(cbind(1, x, i) %*% coef(reg_2), add=TRUE, col=colours[i])
}
```



Based on the figure, nomatter what kind of education levels, mothers should be better to give birth on their late 20s.

Part C

Now create an indicator variable reflecting whether the mother has completed high school or not. Consider interactions between the high school completion and mother's age in family. Also, create a plot that shows the separate regression lines for each high school completion status group.

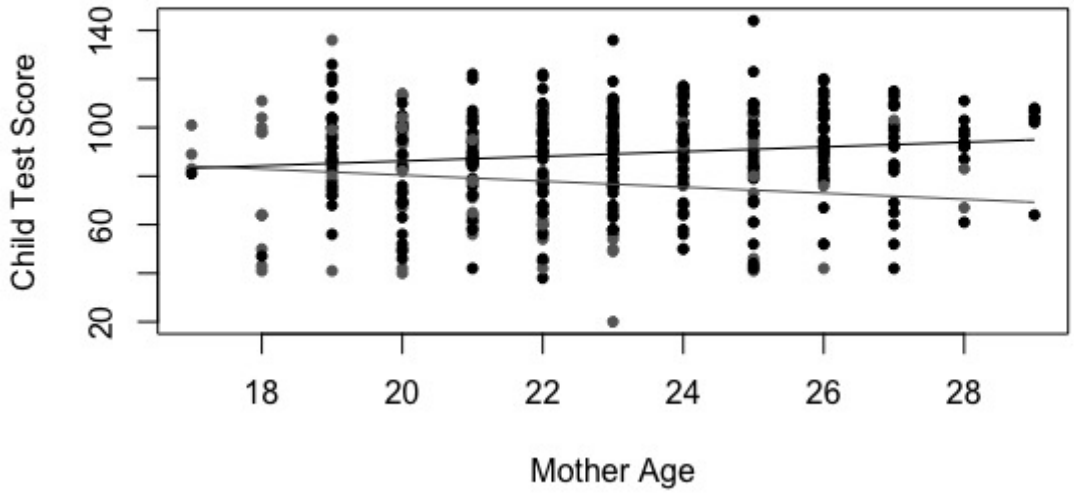
Fit model

```
# binary education level
iq_data$edu_bi <- ifelse(iq_data$educ_cat > 1, 1, 0)
# fit model
reg_3 <- lm(ppvt ~ momage + edu_bi + momage*edu_bi, data=iq_data)
display(reg_3)
```

```
lm(formula = ppvt ~ momage + edu_bi + momage*edu_bi, data = iq_data)
             coef.est coef.se
(Intercept)   105.22    17.65
momage         -1.24     0.81
edu_bi        -38.41    20.28
momage:edu_bi   2.21     0.92
---
n = 400, k = 4
residual sd = 19.85, R-Squared = 0.06
```

Plot

```
colors <- ifelse(iq_data$edu_bi == 1, "#000000", "#666666")
plot(iq_data$momage, iq_data$ppvt, xlab="Mother Age", ylab="Child Test Score", col=co
lors, pch=20)
curve(cbind(1, 1, x, 1 * x) %*% coef(reg_3), add=TRUE, col="#000000")
curve(cbind(1, 0, x, 0 * x) %*% coef(reg_3), add=TRUE, col="#666666")
```



Part D

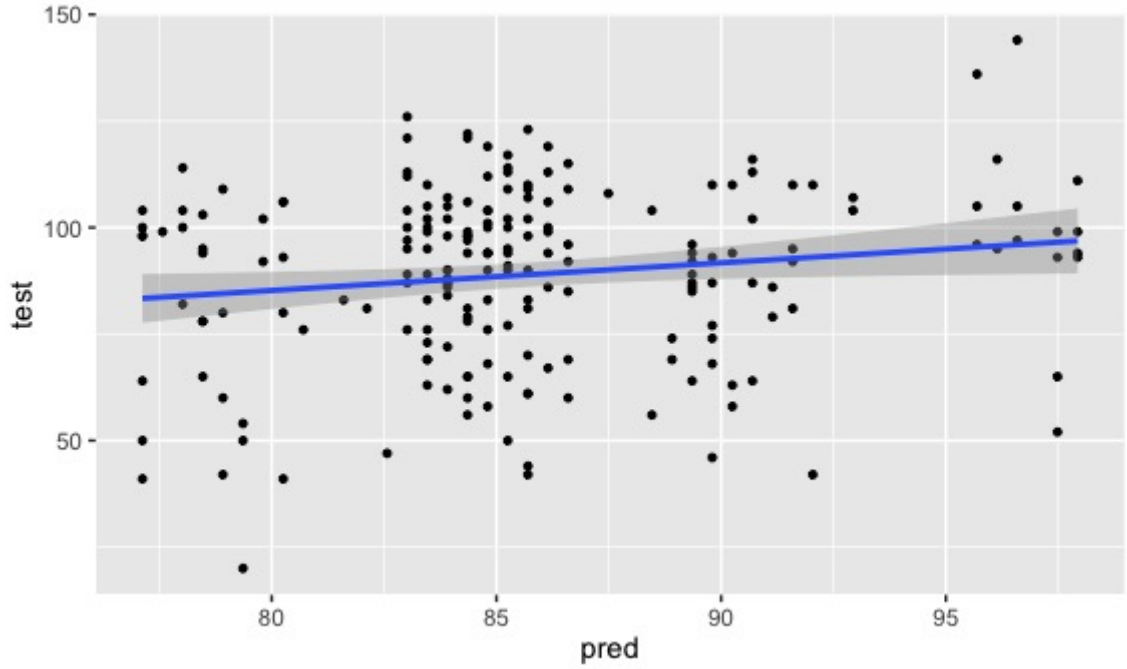
Finally, fit a regression of child test scores on mother's age and education level for the first 200 children and use this model to predict test scores for the next 200. Graphically display comparisons of the predicted and actual scores for the final 200 children.

Fit model

```
# subgroup
iq_train <- iq_data[c(1:200), ]
iq_test <- iq_data[-c(1:200), ]
reg_4 <- lm(ppvt ~ momage + educ_cat, data=iq_train)
display(reg_4)
```

```
lm(formula = ppvt ~ momage + educ_cat, data = iq_train)
            coef.est coef.se
(Intercept) 63.63    11.82
momage       0.45     0.55
educ_cat     5.44     1.82
---
n = 200, k = 3
residual sd = 19.58, R-Squared = 0.06
```

```
# make predictions
iq_pred <- data.frame(predict(reg_4, iq_test))
data_all <- cbind(iq_pred,iq_test$ppvt)
colnames(data_all) <- c('pred','test')
ggplot(data_all,aes(x = pred, y = test)) + geom_point(aes(x = pred, y = test), size =
 1)+
  geom_smooth(method=lm)
```



Based on the plot, this model doesn't predict well in the testing data.

Gelman & Hill 4.4

Logarithmic transformations: the folder pollution contains mortality rates and various environmental factors from 60 U.S. metropolitan areas (see McDonald and Schwing, 1973). For this exercise we shall model mortality rate given nitric oxides, sulfur dioxide, and hydrocarbons as inputs. This model is an extreme oversimplification as it combines all sources of mortality and does not adjust for crucial factors such as age and smoking. We use it to illustrate log transforma- tions in regression.

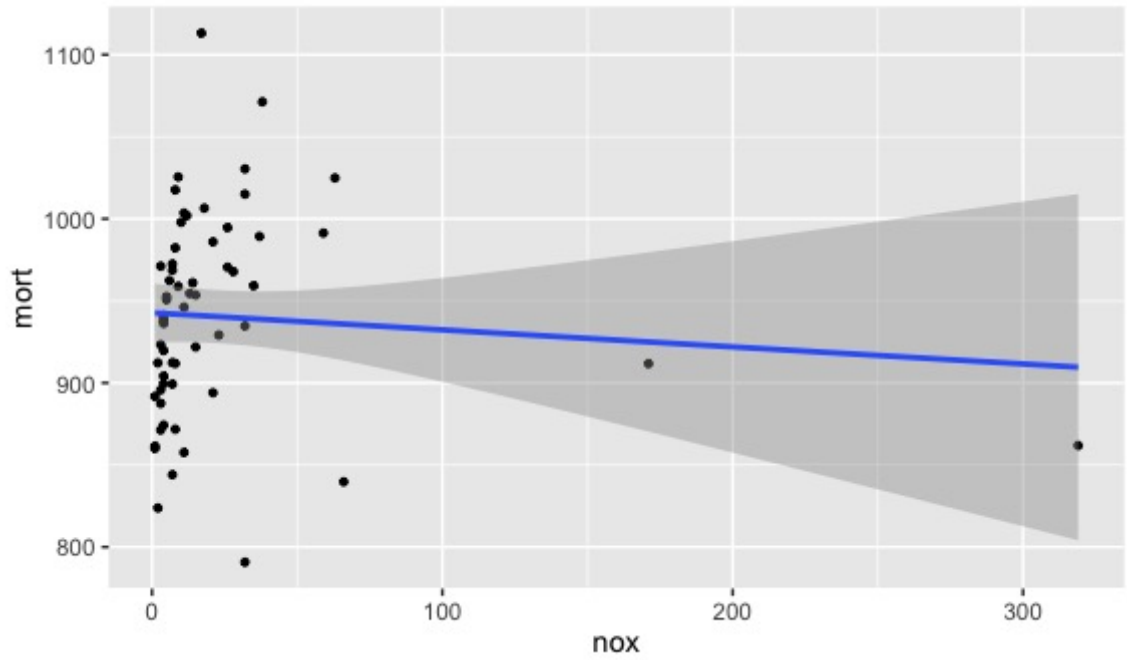
```
pollution <- read.dta ("pollution.dta")
```

Part A

Create a scatterplot of mortality rate versus level of nitric oxides. Do you think linear regression will fit these data well? Fit the regression and evaluate a residual plot from the regression.

Scatterplot

```
ggplot(pollution,aes(x = nox, y = mort)) + geom_point(aes(x = nox, y = mort), size = 1)
 +
  geom_smooth(method=lm)
```



From the scatterplot, I think linear regression will fit these data. However, some outliners may distract the result.

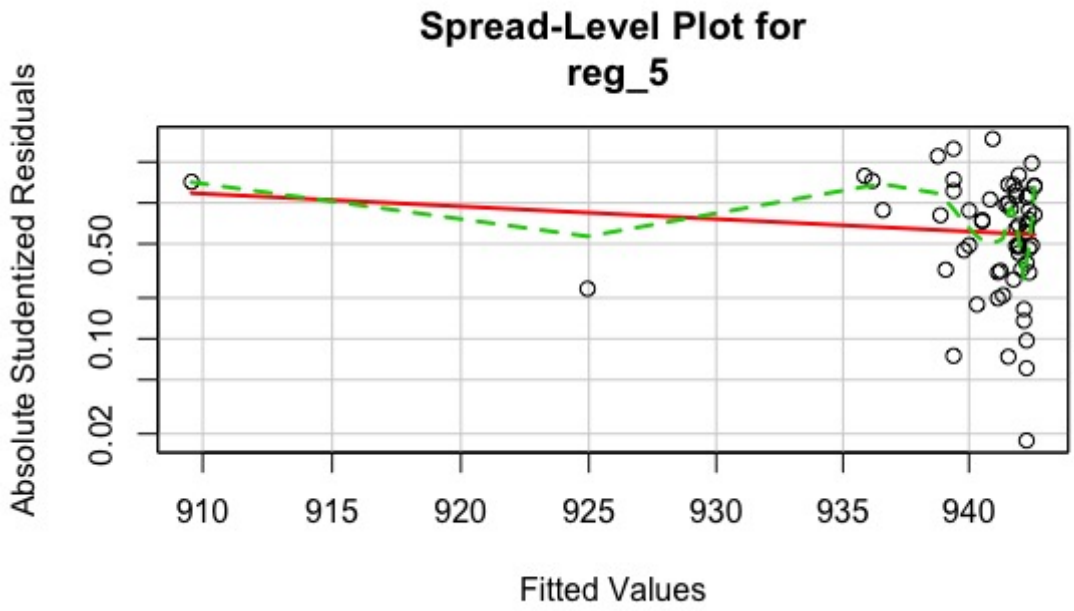
Fit regression model

```
reg_5 <- lm(mort ~ nox, data=pollution)
display(reg_5)
```

```
lm(formula = mort ~ nox, data = pollution)
            coef.est coef.se
(Intercept) 942.71     9.00
nox          -0.10     0.18
---
n = 60, k = 2
residual sd = 62.55, R-Squared = 0.01
```

Residual plot

```
spreadLevelPlot(reg_5)
```



From the residual plot, it can be noted that the model does not fit well as the linearity assumption was not met.

Part B

Find an appropriate transformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.

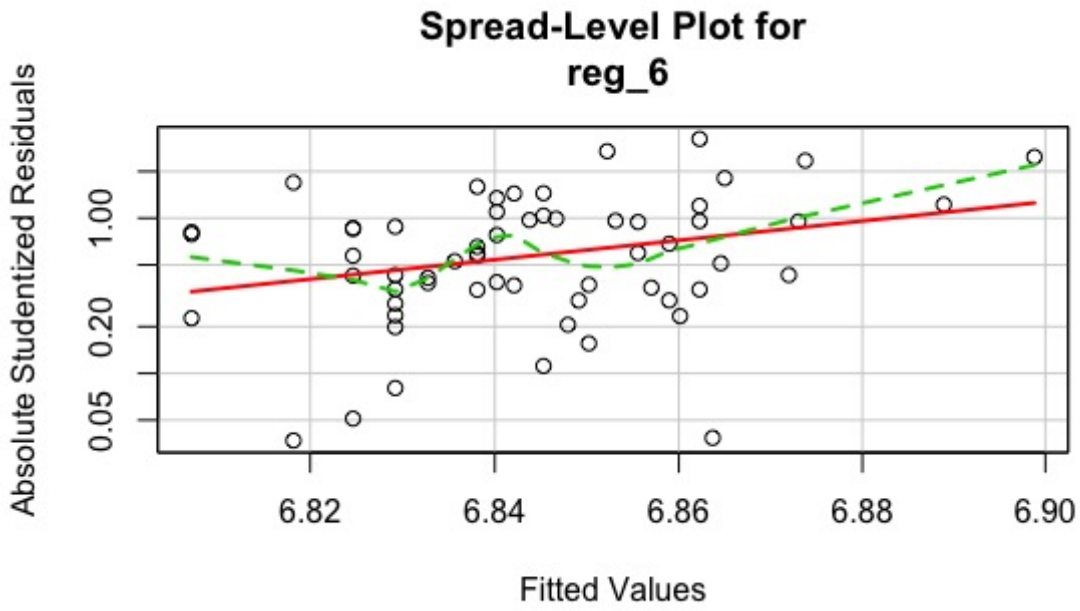
Fit transformation model

```
reg_6 <- lm(log(mort) ~ log(nox), data=pollution)
display(reg_6)
```

```
lm(formula = log(mort) ~ log(nox), data = pollution)
            coef.est coef.se
(Intercept) 6.81     0.02
log(nox)    0.02     0.01
---
n = 60, k = 2
residual sd = 0.06, R-Squared = 0.08
```

Residual plot

```
spreadLevelPlot(reg_6)
```



From the residual plot, it can be noted that the above logarithmic transformations improved the residual plot as the new residuals are more constant as the predicted value increases, as compared with the previous residual plot.

Part C

Interpret the slope coefficient from the model you chose in (b).

Interpretion: for each 1% difference in nitric oxides level, the predicted difference in mortality rate is 0.02%. This association is statistically significant (t=2.255, P=0.0279)

Part D

Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.

logarithmic transformation

```
apply(pollution[, c("hc", "nox", "so2")], FUN=IQR, MARGIN = 2)
```

```
   hc   nox   so2 
23.25 19.75 58.00 
```

```
pollution$log_hc <- log(pollution[, c("hc")])
pollution$log_nox <- log(pollution[, c("nox")])
pollution$log_so2 <- log(pollution[, c("so2")])
```

```
apply(pollution[, c("log_hc", "log_nox", "log_so2")], FUN=IQR, MARGIN = 2)
```

```
  log_hc  log_nox  log_so2 
1.463485 1.779850 1.835902 
```

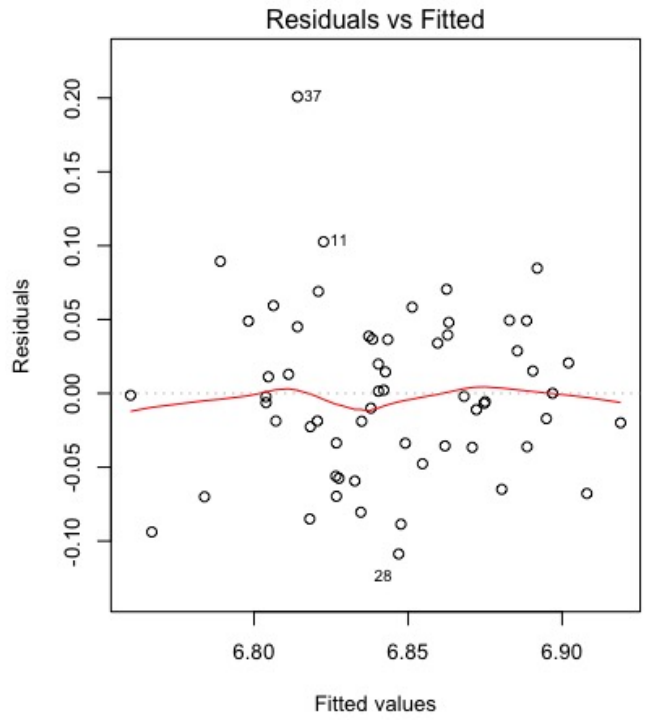
Fit model

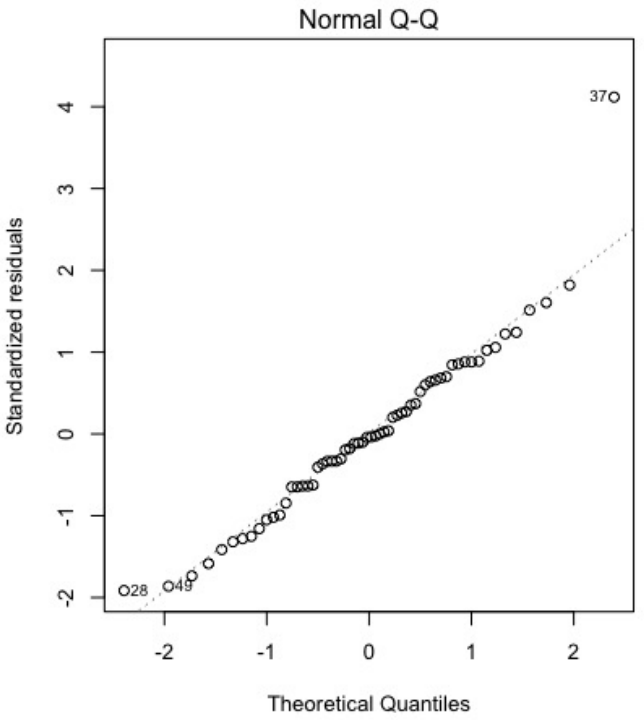
```
reg_7 <- lm(log(mort) ~ log_hc + log_nox + log_so2, data=pollution)
display(reg_7)
```

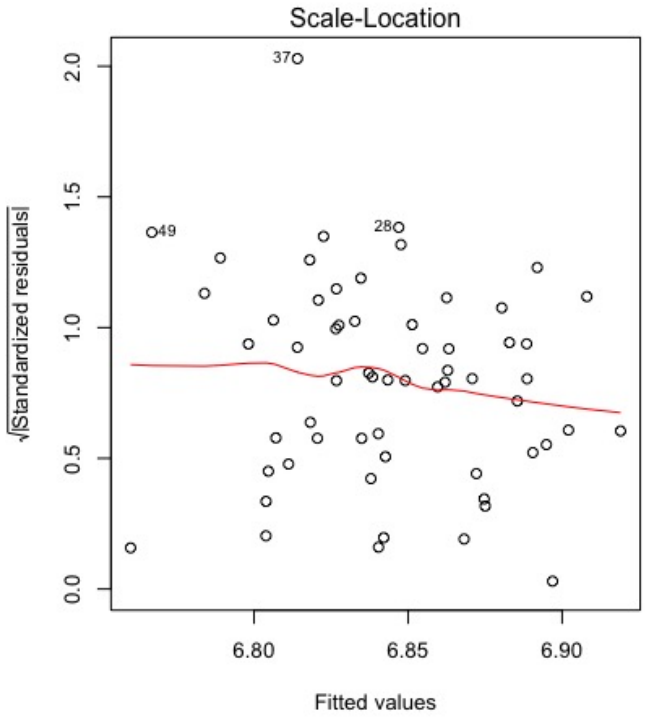
```
lm(formula = log(mort) ~ log_hc + log_nox + log_so2, data = pollution)
            coef.est coef.se
(Intercept)  6.83     0.02
log_hc      -0.06     0.02
log_nox      0.06     0.02
log_so2      0.01     0.01
---
n = 60, k = 4
residual sd = 0.06, R-Squared = 0.29
```

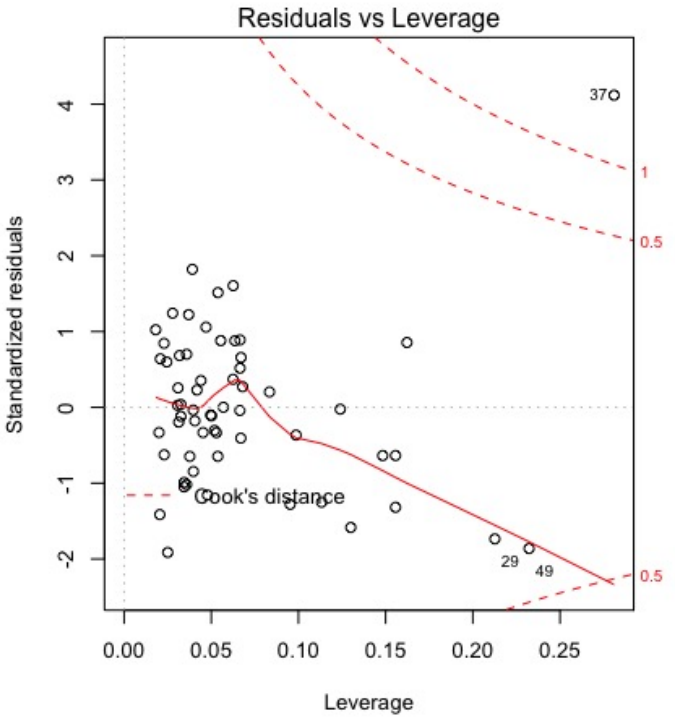
plot the model

```
par(mfrow=c(2,2))
plot(reg_7)
```









Interpretion

Interpretion: for each 1% difference in nitric oxides (NO) level, sulfur dioxide level, and hydrocarbons level, the predicted difference in mortality rate is 0.06%, 0.01%, and -0.06%, respectively.

Part E

Cross-validate: fit the model you chose above to the first half of the data and then predict for the second half. (You used all the data to construct the model in (d), so this is not really cross-validation, but it gives a sense of how the steps of cross-validation can be implemented.)

Fit model

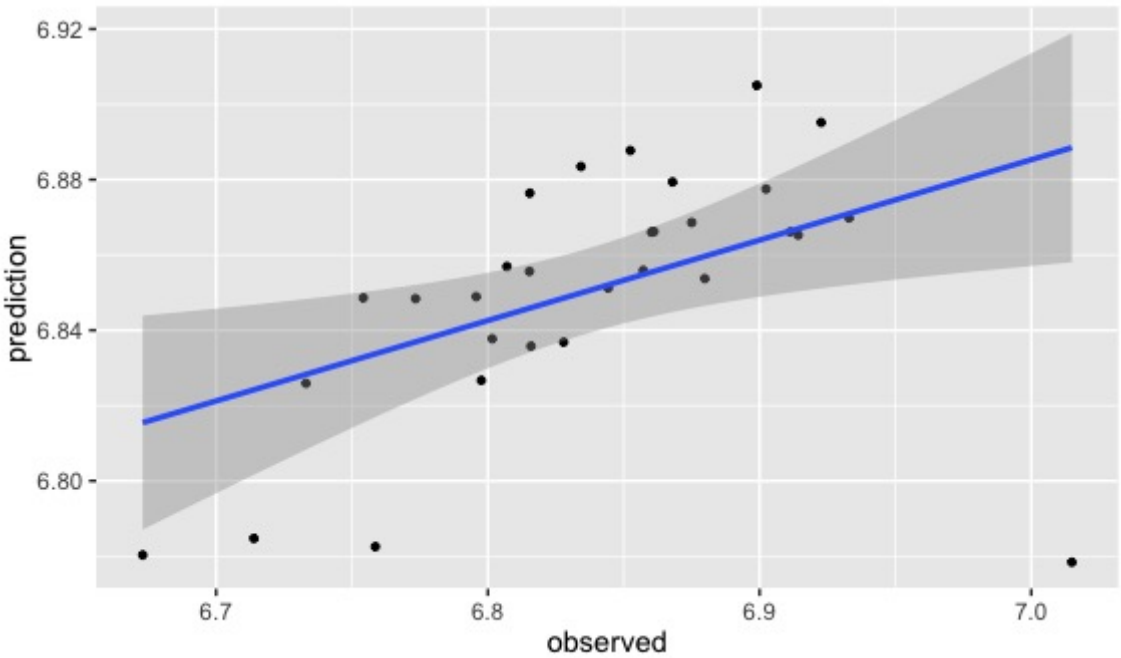
```
# subgroup
training <- pollution[1:30,]
testing <- pollution[31:60,]
reg_8 <- lm(log(mort) ~ log_nox + log_so2 + log_hc, data=training)
display(reg_8)
```

```
lm(formula = log(mort) ~ log_nox + log_so2 + log_hc, data = training)
            coef.est coef.se
(Intercept)  6.80     0.03
log_nox      0.01     0.03
log_so2      0.02     0.01
log_hc      -0.02     0.03
---
n = 30, k = 4
residual sd = 0.06, R-Squared = 0.24
```

Prediction

```
pred <- predict(reg_8, testing)
observ_pred <- data.frame(cbind(log(testing$mort),pred))
colnames(observ_pred) <- c('observed', 'prediction')
```

```
ggplot(observ_pred,aes(x = observed, y = prediction)) + geom_point(aes(x = observed,
y = prediction), size = 1) + geom_smooth(method=lm)
```



Gelman & Hill 5.4

Perform a logistic regression for a problem of interest to you. This can be from a research project, a previous class, or data you download. Choose one variable of interest to be the outcome, which will take on the values 0 and 1 (since you are doing logistic regression).

```
library(mlogit)
library(descr)
# import data
condom <- read.delim('condom.dat', header=T)
str(condom)
```

```
'data.frame':   100 obs. of  7 variables:
 $ safety  : int  3 1 0 3 2 0 0 0 2 3 ...
 $ use     : Factor w/ 2 levels "Condom Used",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ gender  : Factor w/ 2 levels "Female","Male": 1 2 1 2 1 1 2 1 1 2 ...
 $ sexexp  : int  5 3 2 3 3 8 6 5 4 7 ...
 $ previous: Factor w/ 3 levels "Condom used",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ selfcon : int  5 2 3 4 6 5 1 5 0 5 ...
 $ perceive: int  4 2 0 4 3 1 0 1 3 2 ...
 ```
 ```r
condom$safety <- as.numeric(condom$safety)
condom$sexexp <- as.numeric(condom$sexexp)
condom$selfcon <- as.numeric(condom$selfcon)
condom$perceive <- as.numeric(condom$perceive)
```

```
# check IQR
apply(condom[,c(1,4,6,7)], FUN=IQR, MARGIN = 2)
```

```
  safety   sexexp  selfcon perceive
    2.00     4.00     3.25     2.00
```

The differences of IQR between each numeric variables are small. Thus, I didn't make any transformations of the inputs.

Part A and B

Analyze the data in R. Use the display() function to summarize the results. Fit several different versions of your model. Try including different predictors, interactions, and transformations of the inputs.

Fit model

```
#logistic model predicting condom use from perceived risk
condommodel.1<-glm(use ~ perceive, data=condom, family=binomial())
display(condommodel.1)
#logistic model predicting condom use from gender and sex experience
condommodel.2<-glm(use ~ gender + sexexp, data=condom, family=binomial())
display(condommodel.2)
#logistic model with interaction between perceived risk and safety
condommodel.3<-glm(use ~ perceive + safety + perceive*safety, data=condom,
                   family=binomial)
display(condommodel.3)
```

```
glm(formula = use ~ perceive, family = binomial(), data = condom)
            coef.est coef.se
(Intercept)  2.48     0.61
perceive    -0.67     0.16
---
  n = 100, k = 2
  residual deviance = 112.7, null deviance = 136.7 (difference = 23.9)

glm(formula = use ~ gender + sexexp, family = binomial(), data = condom)
            coef.est coef.se
(Intercept)  0.65     0.46
genderMale   0.03     0.41
sexexp      -0.09     0.08
---
  n = 100, k = 3
  residual deviance = 135.3, null deviance = 136.7 (difference = 1.4)

glm(formula = use ~ perceive + safety + perceive * safety, family = binomial,
    data = condom)
                coef.est coef.se
(Intercept)      4.17     1.30
perceive        -1.61     0.43
safety          -0.37     0.40
perceive:safety  0.26     0.11
---
  n = 100, k = 4
  residual deviance = 99.8, null deviance = 136.7 (difference = 36.8)
```

Part C.1

Choose one particular formulation of the model and do the following: Describe how each input affects Pr(y = 1) in the fitted model. You must consider the estimated coefficient, the range of the input values, and the nonlinear inverse-logit function.

```
# final model is formulated as follow
condommodel.2<-glm(use ~ gender + sexexp, data=condom, family=binomial())
exp(cbind(OR = coef(condommodel.2), confint(condommodel.2)))
```

```
                  OR     2.5 %   97.5 %
(Intercept) 1.9067750 0.7888486 4.785707
genderMale  1.0277684 0.4586545 2.298189
sexexp      0.9109053 0.7734751 1.067486
```

The estimate coefficient of genderMale is 0.02739 and the estimate coefficient of sex experience is -0.09332. Since the model has been transformed, the coefficients are not easily interpreted on their own. Instead, the odds ratios are typically calculated and interpreted. The odds of condom use in Male are 1.0277684 times the odds of condom use in Female. This relationship is not statistically significant. The 95% confidence interval indicates the true value of the odds ratio likely lies between 0.7888486 and 4.785707. A one more sex experience increases the odds of condom use 0.9109053 times. The 95% confidence interval indicates the true value of this relationship is between 0.7734751 and 1.067486.

Part C.2

What is the error rate of the fitted model? What is the error rate of the null model?

```
#classification table for percent correctly predicted
p <- predict(condommodel.2, newdata=condom, type="response")
condom$pred <- factor(1*(p >= .5), labels=c("Yes", "No"))
CrossTable(condom$use,condom$pred, expected=F,
           prop.chisq=F, sresid=F, prop.r=T,
           prop.c=F, prop.t=F)
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|-------------------------|

====================================
               condom$pred
condom$use        Yes      No   Total
------------------------------------
Condom Used         9      34      43
               0.209   0.791   0.430
------------------------------------
Unprotected         5      52      57
               0.088   0.912   0.570
------------------------------------
Total              14      86     100
====================================
```

The error rate of the fitted model is (34+5)/100=39%. The error rate of the Null model is 43/100=43%.

Part C.3

Look at the deviance of the fitted and null models. Does the improvement in fit seem to be real?

```
#computing model chi-squared
modelChi <- condommodel.2$null.deviance - condommodel.2$deviance
chidf <- condommodel.2$df.null - condommodel.2$df.residual
chisq.prob <- 1 - pchisq(modelChi, chidf)
modelChi; chidf; chisq.prob
```

```
> modelChi; chidf; chisq.prob
[1] 1.370068
[1] 2
[1] 0.5040732
```

The model was not significantly better than the null model at explaining condom use [χ^2(2)=1.370068; p=0.5040732].

Part C.4

Use the model to make predictions for some test cases of interest.

I used the fitted model to predict the chance of condom use of a female who never have sex before.

Z = 0.64541+0.02739*(genderMale)-0.09332*(sexexp) = 0.62+0.05*0-2.25*0 = 0.62, (e^Z) = e^(0.62) = 1.8589

p(condom use) = (e^Z) / (1 + (e^Z))=1.8589 / (1+1.8589) = 65.02%

The condom use probility of a female who never have sex before is 65.02%.

Gelman & Hill 6.1

Poisson regression: the folder risky.behavior contains data from a random- ized trial targeting couples at high risk of HIV infection. The intervention pro- vided counseling sessions regarding practices that could reduce their likelihood of contracting HIV. Couples were randomized either to a control group, a group in which just the woman participated, or a group in which both members of the couple participated. One of the outcomes examined after three months was “number of unprotected sex acts.”

```
library(AER)
# import data
hiv <- read.dta ("risky_behaviors.dta")
hiv$fupacts <- round(hiv$fupacts)
hiv$couples <- factor(hiv$couples)
hiv$women_alone <- factor(hiv$women_alone)
hiv$bs_hiv <- factor(hiv$bs_hiv, levels=c("negative","positive"))
summary(hiv)
```

```
    sex       couples women_alone      bs_hiv       bupacts          fupacts     
 woman:217   0:272   0:288       negative:337   Min.   :  0.00   Min.   :  0.00  
 man  :217   1:162   1:146       positive: 97   1st Qu.:  5.00   1st Qu.:  0.00  
                                                Median : 15.00   Median :  5.00  
                                                Mean   : 25.91   Mean   : 16.49  
                                                3rd Qu.: 36.00   3rd Qu.: 21.00  
                                                Max.   :300.00   Max.   :200.00  
```

Part A

Model this outcome as a function of treatment assignment using a Poisson regression. Does the model fit well? Is there evidence of overdispersion?

Fit model

```
poi_1 <- glm(fupacts ~ women_alone, family = poisson, data = hiv)
display(poi_1)
```

```
glm(formula = fupacts ~ women_alone, family = poisson, data = hiv)
             coef.est coef.se
(Intercept)   2.92     0.01  
women_alone1 -0.40     0.03  
---
  n = 434, k = 2
  residual deviance = 13064.2, null deviance = 13298.6 (difference = 234.4)
```

The model seems do notfit well, although the woman_alone factor appears to be a statistically significant predictor.

check overdispersion

```
dispersiontest(poi_1,trafo=1)
```

```
    Overdispersion test

data:  poi_1
z = 4.9319, p-value = 4.072e-07
alternative hypothesis: true alpha is greater than 0
sample estimates:
  alpha
41.9765
```

There is evidence of overdispersion (c is estimated to be 41.99409), which speaks quite strongly against the assumption of equidispersion (i.e. c=0).

Part B

Next extend the model to include pre-treatment measures of the outcome and the additional pre-treatment variables included in the dataset. Does the model fit well? Is there evidence of overdispersion?

Fit model

```
poi_2 <- glm(fupacts ~ women_alone + bupacts + bs_hiv, family = quasipoisson, data = hi

display(poi_2)
```

```
glm(formula = fupacts ~ women_alone + bupacts + bs_hiv, family = quasipoisson,
    data = hiv)
             coef.est coef.se
(Intercept)     2.66     0.09
women_alone1   -0.43     0.15
bupacts         0.01     0.00
bs_hivpositive -0.51     0.19
---
  n = 434, k = 4
  residual deviance = 10434.8, null deviance = 13298.6 (difference = 2863.8)
  overdispersion parameter = 29.0
```

check overdispersion

```
dispersiontest(poi_2,trafo=1)
```

```
    Overdispersion test

data:  poi_2
z = 6.6305, p-value = 1.672e-11
alternative hypothesis: true alpha is greater than 0
sample estimates:
   alpha
27.83038
```

There is still evidence of overdispersion (c is estimated to be 27.83038), which speaks quite strongly against the assumption of equidispersion (i.e. c=0).

compare two model

```
anova(poi_1, poi_2, test="Chisq")
```

```
Analysis of Deviance Table

Model 1: fupacts ~ women_alone
Model 2: fupacts ~ women_alone + bupacts + bs_hiv
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       432      13064
2       430      10435  2   2629.3 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Compare to the first model, the second fit of the model has statistically sinificantly improved.

Part C

Fit an overdispersed Poisson model. What do you conclude regarding effectiveness of the intervention?

Fit model

```
poi_3 <- glm(fupacts ~ women_alone + bupacts + bs_hiv, family=quasipoisson, data=hiv)
display(poi_3)
```

```
glm(formula = fupacts ~ women_alone + bupacts + bs_hiv, family = quasipoisson,
    data = hiv)
             coef.est coef.se
(Intercept)     2.66     0.09
women_alone1   -0.43     0.15
bupacts         0.01     0.00
bs_hivpositive -0.51     0.19
---
  n = 434, k = 4
  residual deviance = 10434.8, null deviance = 13298.6 (difference = 2863.8)
  overdispersion parameter = 29.0
```

Based on the results, The number of unprotected sexual acts in the “both-member” group after controlling for baseline HIV infection status and baseline unprotected sexual acts are approximately exp^(-0.43) = 0.65 times that in the “women-alone” group. So the intervention is effective as it reduces 35% of the unsafe sexual acts.

Part D

These data include responses from both men and women from the participating couples. Does this give you any concern with regard to our modeling assumptions?

Yes, this is a problem, because this violates the assumption of data independence of the Possion model.